Building an Updated MEDLINE Co-Occurrences (MRCOC) File

Document Last Updated: Monday, October 06, 2014

Building an Updated MEDLINE Co-Occurrences (MRCOC) File

Table of Contents:

Introduction 2
Acronyms/Abbreviations Used 4
Legacy 2012 MRCOC File 4
Decisions Made in Creating an Updated MRCOC File 5
MeSH Descriptor DUI/CUI Mapping File 6
MeSH Qualifier QUI/CUI Mapping File 6
Special Notes 7
MeSH Indexing File (Indexing.txt) 8
Detailed Descriptor Co-Occurrences File (detailed_CoOccurs_2013.txt) 10
Final MRCOC File (summary_CoOccurs_2013.txt) 11
Example 12
Appendix I: Dates 14

Tables:

Table 1 - Acronyms/Abbreviations Used in This Document 4
Table 2 - Example of XML Translation to ASCII 8

Figures:

Figure 1 - Example from MeSH Descriptor Mapping File 6
Figure 2 - Example from MeSH Qualifier Mapping File 6
Figure 3 - Sample Indexing.txt Data 8
Figure 4 - Sample detailed_CoOccurs_2013.txt Data 10
Figure 5 - Sample summary_CoOccurs_2013.txt Data 11
Figure 6 - Example: Human Indexing to Initial Data File 12
Figure 7 - Detailed Co-Occurrence for Example Citations 13
Figure 8 - Final Co-Occurrence Summary for Example Citations 13

Contact Information:
Author: James G. Mork
Email: jmork@mail.nih.gov
Phone: 301-435-3163

Building an Updated MEDLINE Co-Occurrences (MRCOC) File

Introduction

The MEDLINE co-occurrences file summarizes the MeSH Descriptors that occur together in MEDLINE citations from the MEDLINE/PubMed Baseline[1]. The MEDLINE/PubMed Baseline is a snapshot created at the beginning of each new MeSH Indexing Year[2] containing the MEDLINE, OLDMEDLINE, and PubMed-not-MEDLINE records.

The example to the right shows the indexing from a sample MEDLINE citation on the left and the list of co-occurrences that would be generated from the indexing on the right. We also track whether each of the MeSH Descriptors is considered a Major Topic (starred). In this example, *Poisoning*, *Poisons*, and *Veratrum* are considered Major Topics. A more complete example is available on page 12.

Indexing	Co-Occurrences Generated
MH - *Poisoning MH - *Poisons MH - Vanillic Acid/analogs & derivatives MH - Veratrum/*metabolism	Poisoning — Poisons Poisoning — Vanillic Acid Poisoning — Veratrum Poisons — Vanillic Acid Poisons — Veratrum Vanillic Acid — Veratrum

Asterisks (stars) on MeSH Descriptors and Qualifiers (e.g., Veratrum/*metabolism) designate that they are the Major Topics of the article. Non-Major (non-asterisked) Descriptors and Qualifiers are usually additional topics substantively discussed within the article, terms added to qualify a Major Topic, or Check Tags (excerpt from *Medical Subject Headings (MeSH®) in MEDLINE®/PubMed®: A Tutorial*[3] with slight modification for this description). We specifically identify the co-occurrences where both MeSH Descriptors are marked as Major Topics for backward compatibility with the legacy MRCOC file. The legacy MRCOC file only tracked co-occurrences where both MeSH Descriptors were marked as Major Topics. We now track all MeSH Descriptor co-occurrences for completeness.

For each MEDLINE/PubMed Baseline, we have created two files: One with the complete details (detailed_CoOccurs_YYYY.txt) and then a summarized version (summary_CoOccurs_YYYY.txt) of the identified co-occurrences. The summary file is the replacement for the legacy UMLS MRCOC file. The more detailed file provides deeper and richer data if the information is required. For example, from the detailed file, you can identify all of the PMIDs where the MeSH Descriptors *Poisoning* and *Vanillic Acid* co-occur and are both identified as Major Topics. Or, if you are trying to identify the earliest paper talking about both *Poisoning* and *Vanillic Acid*, you will find the information in the detailed file.

The UMLS (Unified Medical Language System) MRCOC file has historically tracked the co-occurrences of important concepts from three sources: MEDLINE, AI/RHEUM (The Artificial Intelligence Rheumatology Consultant System), and CCPSS (The Canonical Clinical Problem Statement System). Co-occurrences in this context are concepts that occur together within a single information source like a MEDLINE Citation or an AI/RHEUM record.

The MRCOC file was created and distributed by the UMLS group[4] as part of their regular releases up through the 2013AA release. The MRCOC file provided by the UMLS group included the following co-occurrence information:

- From MEDLINE, co-occurrence data were computed for concepts that were designated as principal or main points in the same journal article, i.e., the co-occurrence counts do not include articles in which either or both of the concepts were present and indexed in MEDLINE but not designated as main points. Two overall frequencies of MEDLINE co-occurrence were provided: one for recent MEDLINE data (MED) and one for MEDLINE data from a preceding block of years (MBD).
- The AI/RHEUM co-occurrence data represented the co-occurrence of diseases and findings in the AI/RHEUM knowledge base, i.e., the diseases that co-occur with a particular finding and the findings that co-occur with a particular disease. Each disease/finding pair can co-occur only once in the AI/RHEUM knowledge base.
- In CCPSS, the co-occurrence data were extracted from patient records and included problem-problem co-occurrences within a patient record as well as problem-modifier co-occurrences.

The decision was made in late 2012 that the UMLS group would no longer be creating and distributing the MRCOC file as part of their normal releases. There were several reasons for this decision including:

[1] http://www.nlm.nih.gov/bsd/licensee/baseline.html
[2] http://www.nlm.nih.gov/bsd/mesh_indexing_date_range.html
[3] http://www.nlm.nih.gov/bsd/disted/meshtutorial/principlesofmedlinesubjectindexing/majortopics/
[4] http://www.nlm.nih.gov/research/umls/

1. The file really did not fit with the rest of the data that the UMLS provides.
2. There was no reason to include this information behind the Metathesaurus License restrictions.
3. The data in the MRCOC file never coincided with the actual MEDLINE Baseline it was meant to represent due to timing constraints of when it needed to be created and when the actual MEDLINE Baseline was created. So, there was always a slight disconnect between the MRCOC file and the Baselines.

In this late 2012 meeting, it was also decided that the Cognitive Science Branch (CgSB)[5] of the Lister Hill National Center for Biomedical Communications, an Intramural Research Division of the U.S. National Library of Medicine[6] would take over the creation and distribution of the MRCOC file. The MRCOC file is a critical component of the research being done in the CgSB branch and personnel in the branch are already working with the yearly MEDLINE Baseline to produce statistics[7], making it a natural fit. With this decision came the opportunity to make improvements to the MRCOC file adding information that might expand the usability and to remove outdated material. The *Decisions Made in Creating an Updated MRCOC File* section details the modification decisions that were made in producing this updated version of the MRCOC file.

The main updates to the MRCOC file were to focus solely on the MEDLINE Baseline information and identify all of the MeSH Heading co-occurrences while still providing the ability to identify the same MEDLINE related information that was included in the legacy MRCOC files. The AI/RHEUM and CCPSS data are not available to update the information for use in the new MRCOC file. The existing AI/RHEUM and CCPSS records from the MRCOC file are available in the historical versions of the UMLS releases up through the 2013AA release. The MRCOC web page[8] also contains a static file with the 2013AA version of the AI/RHEUM and CCPSS data.

The 2013 Baseline Year is the first iteration of the updated MRCOC file; please send your questions, comments, and enhancement suggestions for either the file or documentation to *metamap@nlm.nih.gov*.

[5] http://lhncbc.nlm.nih.gov/branch/cognitive-science-branch
[6] http://lhncbc.nlm.nih.gov/
[7] http://www.nlm.nih.gov/bsd/licensee/baselinestats.html
[8] http://mbr.nlm.nih.gov/MRCOC.shtml

Acronyms/Abbreviations Used

Acronym	Description
AI/RHEUM	The Artificial Intelligence Rheumatology Consultant System (legacy 2012 UML MRCOC)
AUI	UMLS Atom Unique Identifier (legacy 2012 UML MRCOC)
CCPSS	The Canonical Clinical Problem Statement System
COA	Attributes of co-occurrence (legacy 2012 UML MRCOC)
COF	Frequency of co-occurrence (legacy 2012 UML MRCOC)
COT	Type of co-occurrence (legacy 2012 UML MRCOC)
CUI	UMLS Concept Unique Identifier
DUI	MeSH Descriptor Unique Identifier
KN	Negative association in Knowledge Base, e.g., a finding that is inconsistent with a disease [AI/Rheum] (legacy 2012 UML MRCOC)
KP	Positive association in Knowledge Base [AI/Rheum] (legacy 2012 UML MRCOC)
L	Co-occurrence of primary or main subject headings in citations to the published literature (legacy 2012 UML MRCOC)
LQ	Second concept occurs as a MeSH topical qualifier of the first in citations to the published literature (legacy 2012 UML MRCOC)
LQB	Second concept is qualified by the first (a MeSH topical qualifier) in citations to the published literature (legacy 2012 UML MRCOC).
MED	MEDLINE data from the most recent five years (years 1 – 5 from generated year) [2012-2008 for 2013]
MBD	MEDLINE data from a preceding block of five years (years 6-10 from generated year) [2007 – 2003 for 2013]
MeSH	Medical Subject Headings
MP	Co-occurrence of modifier and problem within a patient record [CCPSS] (legacy 2012 UML MRCOC)
MRCOC	UMLS Co-occurrences file distributed with each UMLS Metathesaurus release
PP	Co-occurrence of two problems within a patient record [CCPSS] (legacy 2012 UML MRCOC)
QUI	MeSH Qualifier Unique Identifier
RST	MEDLINE data from prior to MBD block of years (years 11 back to the beginning) [2002 – 1965 for 2013]
SAB	Abbreviation of the source of co-occurrence information (legacy 2012 UML MRCOC)
UMLS	Unified Medical Language System
ZN	Flag added to new MRCOC files denoting whether both main headings are Major Topics (ZN – No)
ZY	Flag added to new MRCOC files denoting whether both main headings are Major Topics (ZY – Yes)

Table 1 - Acronyms/Abbreviations Used in This Document

Legacy 2012 MRCOC File

- Includes UMLS CUIs and AUIs
- Only includes co-occurrences where both CUI1 and CUI2 are designated as Major Topics (Main Points) of the citation.
- Tracks co-occurrences in both directions (CUI1:CUI2 and CUI2:CUI1) with the MeSH Qualifiers assigned to whichever CUI is on the left in the pairing.
- Uses the latest date for the citation, either Last Revised or Date Completed to determine MED (last 5 years) and MBD (5 years prior to MED) status denoted in the SAB (Source of Co-occurrence Abbreviation) field.
- Includes AI/RHEUM and CCPSS data denoted by “KP”, “KN”, “MP”, and “PP” in the COT (Type of Co-occurrence) field.
- Includes MeSH Qualifier to MeSH Descriptor co-occurrences denoted by “LQ” and “LQB” in the COT field.
- Includes MeSH Descriptor co-occurrences denoted by “L” in the COT field.

Building an Updated MEDLINE Co-Occurrences (MRCOC) File

Decisions Made in Creating an Updated MRCOC File

Several decisions were made when the topic of creating a new version of the MRCOC file was discussed. While it was agreed that the information in the MRCOC file was important, it was felt that we could improve the usability by making some changes.

Removals:

- The UMLS AUI (Unique Identifier of first atom) information is not included.
- The Content View Flag (CVF) field is not included.
- There are no MeSH Qualifier co-occurring relationships (LQ & LQB).
- The AI/RHEUM and CCPSS data are not included in future MRCOC files. The data are outdated and we have no means of recalculating at this point. In the 2012AB version of the MRCOC file, MEDLINE accounts for 99.65% of the data.

Additions/Changes:

- The new MRCOC files are created using the latest MEDLINE Baseline (2013) shortly after release of the Baseline.
- We added the MeSH DUI (Descriptor Unique Identifier) and MeSH QUI (Qualifier Unique Identifier) information to the new files.
- The new MRCOC includes information on all co-occurring MeSH Descriptors and not just where the two MeSH Descriptors were considered Major Topics (starred) of the citation. A new flag indicates whether both were starred (*ZY*) or not (*ZN*).
- We provide the ability to easily identify PMIDs associated with each co-occurring pair of MeSH Descriptors and Qualifier/Descriptor co-occurrences.
- We added multiple dates to the records: Publication Date, Article Date, Date Completed, MeSH Indexing Year, and computing the earliest of these dates. See **Appendix I** for full descriptions and examples of these date fields and a list of defaults used in the conversion process for Publication Date. For example, if Publication Date is missing either the month or day field, we assign the value "1".
- We created co-occurrences in only one direction. The new MRCOC file only includes the co-occurrence in a single direction, and this single relationship contains counts for all Qualifiers associated with both Descriptors. Co-occurrences are still searchable in either direction, but, only a single entry appears in the data. See "Example" section for a full description and examples of this.
- We use the MEDLINE Baseline Year minus the Year from the Date Completed date to consistently compute "recent MEDLINE" (MED), "preceding block of MEDLINE" (MBD), and the rest of MEDLINE (RST). We include yearly summarized sets of frequency counts for each co-occurrence, while a year is only included if the co-occurring frequency is greater than zero for that year!
    - **MED:** All occurrences in the last 5 years of the MEDLINE Baseline for each set of co-occurring MeSH Descriptors. For the 2013 MEDLINE Baseline, the last 5 years would include 2008 – 2012.
    - **MBD:** All occurrences in the 5 years prior to the **MED** set (years 6 - 10) of the MEDLINE Baseline for each set of co-occurring MeSH Descriptors. For the 2013 MEDLINE Baseline, years 6 - 10 would include 2003 - 2007.
    - **RST:** All occurrences in the years prior to the **MBD** set (years 11 backwards to the beginning) of the MEDLINE Baseline for each set of co-occurring MeSH Descriptors. For the 2013 MEDLINE Baseline, this would include 2002 – 1965.

MeSH Descriptor DUI/CUI Mapping File

We used the 2012AB UMLS Metathesaurus version of the ***MRCONSO.RRF*** file to identify the UMLS CUIs (column 1) and Concept Names (column 15) associated with each of the 2013 MeSH Descriptors and combined that with the DUIs (UI) from the MeSH ***d2013.bin*** file to create a single file allowing us to go from MeSH Descriptor ➢ CUI ➢ DUI. This combined file is used to map the full text of the MeSH Descriptors found in the individual MEDLINE Baseline XML files into DUIs and CUIs. The following Unix command was used to identify the UMLS CUIs for each MeSH Descriptor.

```
grep '|MSH|MH|' MRCONSO.RRF | grep "|ENG|" | cut -d'|' -f1,15 > MHcui
```

The combined file as shown in Figure 1 is a bar "|" separated flat file containing the following fields:

- **UMLS CUI** – UMLS Concept Unique Identifier for Descriptor
- **MeSH DUI** – Descriptor Unique Identifier for Descriptor
- **MeSH Descriptor** – Preferred name of the MeSH Descriptor

```
C0242904|D018686|Anesthetics, Intravenous
```

Figure 1 - Example from MeSH Descriptor Mapping File

MeSH Qualifier QUI/CUI Mapping File

We used the 2012AB UMLS Metathesaurus version of the ***MRCONSO.RRF*** file to identify the UMLS CUIs (column 1) and Concept Names (column 15) associated with each of the 2013 MeSH Qualifiers and combined that with the QUIs (UI) and Two-Character Abbreviations (QA) from the MeSH ***q2013.bin*** file to create a single file allowing us to go from MeSH Qualifier ➢ CUI ➢ QUI. This combined file is used to map the full text of the MeSH Qualifiers found in the individual MEDLINE Baseline XML files into QUIs, CUIs, and their two-character abbreviations. The following Unix command was used to identify the UMLS CUIs for each MeSH Qualifier.

```
grep '|MSH|TQ|' MRCONSO.RRF | grep "|ENG|" | cut -d'|' -f1,15 > SHcui
```

The combined file as shown in Figure 2 is a bar "|" separated flat file containing the following fields:

- **UMLS CUI** – UMLS Concept Unique Identifier for Descriptor
- **MeSH QUI** – Qualifier Unique Identifier for Descriptor
- **MeSH Qualifier** – Preferred name of the MeSH Qualifier
- **Two-Char Abbreviation** – Two-Character Abbreviation for the MeSH Qualifier

```
C0023242|Q000331|legislation & jurisprudence|LJ
```

Figure 2 - Example from MeSH Qualifier Mapping File

There are no plans to distribute either of these mapping files. The files are described here to explain how and where the data is being generated for the final summarized MRCOC file.

Special Notes

- In this document, we use the terms *article* and *citation* interchangeably, but they do refer to two distinct entities in the indexing world. Indexers index from the full text of an *article* and the results of that effort along with the title, abstract, and other bibliographic information from the *article* are stored as a *citation* in the MEDLINE/PubMed database.
- All indexed citations in the 2013 MEDLINE Baseline are Version 1. Although the field is available for referencing different versions of the same PMID, in this baseline, none of the indexed citations have anything other than version "1" assigned.
- MeSH Descriptors are the main descriptors or headings from the MeSH Vocabulary (e.g., *Lung*) used to describe the major topics of an *article*. MeSH Qualifiers or subheadings are used to qualify the MeSH Descriptors (e.g., *Lung/abnormalities* means that the article is more about the *abnormalities* associated with the *Lung* than the *Lung* itself).
- If any single MeSH Qualifier associated with a MeSH Descriptor is starred (marked as a Major Topic), the MeSH Descriptor will inherit that designation. A MeSH Descriptor can also be starred on its own. In the event that a starred Qualifier is removed from a Descriptor, that Descriptor will inherit the starred designation.
- The situation where the star is on the MeSH Descriptor and not on the MeSH Qualifier is called an *Upfront Major Topic* and is relatively rare.
- Descriptor/Qualifier co-occurrences may contain what would be considered non-allowed combinations based on the current (2013) version of MeSH. The reason for this is historical. A Descriptor has a list of allowable Qualifiers and that list can change over time. The Qualifiers are not removed when a Descriptor is updated or when the Qualifier is no longer allowed. For example, PMID: 2614180 has the following "*MH - Receptors, Transferrin/*analysis/isolation & purification/pharmacology*". If we look at the MeSH Record for Descriptor "*Receptors, Transferrin*", we can see that "*pharmacology*" is no longer an allowed Qualifier. Sometime between when this citation was indexed on 19900314 and release of the 2000 MeSH (the earliest version of MeSH we maintain), the Qualifier was dropped from the list of allowable Qualifiers for this Descriptor. So, these non-allowed combinations will show up in the detailed and summary files.

MeSH Indexing File (Indexing.txt)

We used the 2013 MEDLINE Baseline set of XML files to create a single file (Indexing.txt) incorporating all of the MeSH Descriptors and associated MeSH Qualifier information for each citation. All XML tags associated with the MeSH Descriptors and Qualifiers are converted to their ASCII equivalent. Table 2 contains an example of this conversion. This file was created simply to summarize the indexing information so we don't have to go back through all of the MEDLINE Baseline files if we have to rerun.

XML Tagging	ASCII Equivalent
<DescriptorName MajorTopicYN="N"> Anesthetics, Intravenous </DescriptorName>	Anesthetics, Intravenous
<QualifierName MajorTopicYN="Y">legislation & jurisprudence</QualifierName>	legislation & jurisprudence

Table 2 - Example of XML Translation to ASCII

The Indexing.txt file contains one line for every MeSH Descriptor that was manually indexed for each citation found in the 2013 MEDLINE Baseline with at least one MeSH Descriptor. Information about the MeSH Descriptor, any associated MeSH Qualifiers, whether the MeSH Descriptor is considered a Major Topic of the citation, and whether or not each of the MeSH Qualifiers is considered a Major Topic of the citation is tracked. Just a reminder that a MeSH Descriptor inherits the Major Topic status of its associated MeSH Qualifiers, so if even only one Qualifier is considered a Major Topic, the MeSH Descriptor will be considered a Major Topic. The MeSH Descriptors and Qualifiers are included in the ordering that they appear in the indexing[9]. We consider this file to just be a summarization step in the process of building the updated MRCOC file. The Indexing.txt file as shown in Figure 3 is a bar "|" separated flat file containing the following information:

- **PMID** – PubMed Unique Identifier
- **Version** – PMID Version number
- **Earliest Date** – Earliest computed date between PubDate, Article Date, and Date Completed
- **Publication Date** – Date article was published (YYYYMMDD)
- **Article Date** – Electronically Published Date (YYYYMMDD), is zero (0) if field is not present in the citation
- **Date Completed** – Date Completed information (YYYYMMDD)
- **MeSH Indexing Year** – Year computed from the Date Completed date (YYYY)
- **Major Topic?** – Inherited or on its own - Flag (1 - Yes, 0 - No)
- **Upfront Major Topic?** – Only set if Descriptor itself is a Main Topic - Flag (1 - Yes, 0 - No)
- **UMLS CUI** – UMLS Concept Unique Identifier for Descriptor
- **MeSH DUI** – Descriptor Unique Identifier for Descriptor
- **Number of Unique Qualifiers** assigned to Descriptor
- **List of Qualifiers** – Comma separated triplet list (if any) of Qualifiers assigned to Descriptor
  The triplet consists of three colon ":" separated fields (e.g., 1:PD:Q000494):
  - **Qualifier Main Point?** – Flag (1 – Yes, 0 – No)
  - **Qualifier Abbreviation**
  - **Qualifier Unique Identifier**

```
20278133|1|19461001|19461001|0|20100318|2010|1|1|C0003855|D001164|0||
19928636|1|19280401|19280401|0|20091231|2010|0|0|C0020405|D006920|3|0:EC:Q000191,0:ED:Q000193,0:HI:Q000266|
22958912|1|20120905|20121027|20120905|20121108|2012|1|0|C0018790|D006323|2|1:MO:Q000401,0:TH:Q000628|
```

Figure 3 - Sample Indexing.txt Data

In Figure 3, reading from left to right - the first entry is for Version 1 of PMID 20278133 where the earliest date and Publication Date are 19461001, there was no Article Date, Date Completed is 20100318, and the MeSH Indexing year is 2010. The MeSH Descriptor *Arteriovenous Fistula* is a Major Topic, has CUI "C0003855", DUI "D001164", and there were no MeSH Qualifiers assigned to the Descriptor. The second entry is for Version 1 of PMID 19928636 where the earliest date and Publication Date are 19280401, there was no Article Date, Date Completed is 20091231, and the MeSH Indexing year is 2010. The MeSH Descriptor *Hygiene* was not a Major Topic, has CUI "C0020405", DUI "D006920", and has three assigned MeSH Qualifiers: *economics* (EC) is not a Major Topic and has QUI Q000191, *education* (ED) is not a Major Topic and has QUI Q000193, and *history* (HI) is not a Major Topic and has QUI Q000266. The third entry is for Version 1 of PMID 22958912 where the earliest date is 20120905, the Publication Date is 20121027, the Article Date is 20120905, Date Completed is 20121108, and the MeSH Indexing year is 2012. The MeSH

[9] Currently this is alphabetical order for both Descriptors and Qualifiers

Descriptor *Heart Arrest* is an inherited Major Topic, has CUI "C0018790", DUI "D006323", and has two assigned MeSH Qualifiers: *mortality* (MO) is a Major Topic and has QUI Q000401 and *therapy* (TH) is not a Major Topic and has QUI Q000628.

There are no plans to distribute the Indexing.txt data file. The file is described here to explain how and where the data is being generated for the final summarized MRCOC file. This file is also large at 15 GB uncompressed.

Detailed Descriptor Co-Occurrences File (detailed_CoOccurs_2013.txt)

The detailed descriptor co-occurrences file contains the complete information for each MeSH Descriptor co-occurrence and allows for identifying PMIDs for specific sets of co-occurrences. The file is sorted into DUI1, DUI2, Completed Year, and PMID order clustering all of the DUI1/DUI2 co-occurrence combinations by the year completed for easier summarization. The file also contains information identifying which MeSH Qualifiers are associated with which MeSH Descriptors in the co-occurrence. So, it is possible to replicate the legacy MRCOC file *LQ* and *LQB* two-way view from this file if desired. The file as shown in Figure 4 is a bar "|" separated flat file containing the following information:

- **PMID** – PubMed Unique Identifier
- **Version** – PMID Version number
- **Earliest Year** – Earliest computed date between PubDate, Article Date, and Date Completed
- **PubDate** – Date article was published (YYYYMMDD)
- **Article Date** – Electronically Published Date (YYYYMMDD), is zero (0) if field is not present in the citation
- **Date Completed** – Date Completed information (YYYYMMDD)
- **MeSH Indexing Year** – The MeSH Indexing year computed from the Date Completed date. Typical current MeSH Indexing years run from mid-November to mid-November the following year. For example, MeSH Indexing Year 2012 ran from November 18, 2011 to November 13, 2012.
- **Both Major Topics** – Flag (ZY – Yes, ZN – No)
- **Information for MeSH Descriptor 1**:
  - **MeSH DUI** – Descriptor Unique Identifier for Descriptor 1
  - **Major Topic?** – Inherited or on its own - Flag (1 – Yes, 0 – No)
  - **Upfront Major Topic?** – Only set if Descriptor itself is a Main Topic - Flag (1 - Yes, 0 - No)
  - **UMLS CUI** – UMLS Concept Unique Identifier for Descriptor 1
  - **Number of Unique Qualifiers** assigned to Descriptor 1
  - **List of Qualifiers** – Comma separated triplet list (if any) of Qualifiers assigned to Descriptor 1. The triplet consists of three colon ":" separated fields (e.g., 1:PD:Q000494):
    - **Qualifier Main Point?** – Flag (1 – Yes, 0 – No)
    - **Qualifier Abbreviation**
    - **Qualifier Unique Identifier**
- **Information for MeSH Descriptor 2**:
  - **MeSH DUI** – Descriptor Unique Identifier for Descriptor 2
  - **Major Topic?** – Inherited or on its own - Flag (1 – Yes, 0 – No)
  - **Upfront Major Topic?** – Only set if Descriptor itself is a Main Topic - Flag (1 - Yes, 0 - No)
  - **UMLS CUI** – UMLS Concept Unique Identifier for Descriptor 2
  - **Number of Unique Qualifiers** assigned to Descriptor 2
  - **List of Qualifiers** – Comma separated triplet list (if any) of Qualifiers assigned to Descriptor 2. The triplet consists of three colon ":" separated fields (e.g., 0:ME:Q000378):
    - **Qualifier Main Point?** – Flag (1 – Yes, 0 – No)
    - **Qualifier Abbreviation**
    - **Qualifier Unique Identifier**

```
16094961|1|19670901|19670901|0|20050923|2005|ZN|D003731|1|0|C0011334|1|1:PA:Q000473|D006801|0|0|C0086418|0||
20278133|1|19461001|19461001|0|20100318|2010|ZY|D001808|1|0|C0005847|1|1:AB:Q000002|D006225|1|0|C0018563|1|1:BS:Q000098|
```

Figure 4 - Sample detailed_CoOccurs_2013.txt Data

In Figure 4, the first entry is for Version 1 of PMID 16094961, where the earliest date and Publication Date are 19670901, there was no Article Date, Date Completed is 20050923, and the MeSH Indexing year is 2005. The two DUIs were both not Major Topics (ZN), DUI1 *Dental Caries* (D003731) is an inherited Major Topic, has CUI1 (C0011334), and has one MeSH Qualifier *pathology* (PA/Q000473) assigned which is a Major Topic. DUI2 *Humeral Fractures* (D006801) was not a Major Topic, has CUI2 (C0086418), and had no MeSH Qualifiers assigned. The second entry shows similar information, except that both DUI1 and DUI2 were Major Topics (ZY) and both had one MeSH Qualifier assigned.

The **detailed_CoOccurs_2013.txt** file will be made available for download. The file is fairly large (125 GB uncompressed, 16 GB compressed), but provides users with the ability to identify all of the PMIDs associated with any co-occurring pairing by a simple grep command "***grep D003731 detailed_CoOccurs_2013.txt | grep D006801***" which allows the DUIs to be specified in any order.

Final MRCOC File (summary_CoOccurs_2013.txt)

Since the ***detailed_CoOccurs_2013.txt*** file is already sorted into order by DUI1/DUI2/Year/PMID it is a simple matter to summarize all of the co-occurrences and calculate frequency counts. The co-occurrence frequencies are summarized by Date Completed Year and assigned the appropriate TimeFrame or SAB designation based on the Date Completed Year. Entries for a given co-occurrence pair are only included in the file when the frequency count is greater than zero, so there may be interruptions in the years based on the frequency of the pairing. The file as shown in Figure 5 is a bar "|" separated flat file containing the following information:

- **MeSH DUI1** – Descriptor Unique Identifier for Descriptor 1
- **UMLS CUI1** – UMLS Concept Unique Identifier for Descriptor 1
- **MeSH DUI2** – Descriptor Unique Identifier for Descriptor 2
- **UMLS CUI2** – UMLS Concept Unique Identifier for Descriptor 2
- **Overall Frequency** – Overall Frequency count for DUI1/DUI2 co-occurrence
- **Starred Frequency (COF)** – Frequency count for DUI1/DUI2 co-occurrence when both DUI1 and DUI2 are considered Major Topics (inherited or naturally)
- **Date Completed Year** – Year from the Date Completed date.
- **TimeFrame (SAB)** – Abbreviation for the timeframe (RST, MBD, MED) of the co-occurrence information. MED refers to the most current five (5) years of MEDLINE (years 1-5). MBD refers to the five years preceding MED (years 6-10). RST refers to entries that fall outside of the most current 10 years of MEDLINE.
- **Frequency No Qualifiers** – Number of occurrences where neither DUI1 or DUI2 had any Qualifiers assigned
- **Frequency Starred No Qualifier** – Number of occurrences where neither DUI1 or DUI2 had any Qualifiers assigned when both are Major Topics
- **Frequency Starred Qualifiers** – Number of occurrences when Qualifiers were assigned to both DUI1 and DUI2 when both DUI1 and DUI2 are Major Topics
- **Both Major Topics** – Flag denoting when Starred Frequency (COF) > 0, or not (ZY – Yes, ZN – No)
- **Frequency DUI1 Only Starred** – Number of occurrences DUI1 is a Major Topic and DUI2 is not.
- **Frequency DUI2 Only Starred** – Number of occurrences DUI2 is a Major Topic and DUI1 is not.
- **Frequency DUI1 Only has Qualifier –** Number of occurrences where DUI1 has Qualifiers assigned, but, DUI2 doesn't.
- **Frequency DUI2 Only has Qualifier –** Number of occurrences where DUI2 has Qualifiers assigned, but, DUI1 doesn't.

```
D000001|C0000699|D000515|C0002191|1|0|1989|RST|0|0|1|ZN|0|0|0|0
D000001|C0000699|D000515|C0002191|1|0|1990|RST|0|0|1|ZN|1|0|0|0
D000001|C0000699|D000515|C0002191|1|0|1993|RST|0|0|1|ZN|0|0|0|0
D000001|C0000699|D000515|C0002191|1|1|1995|RST|0|0|1|ZY|0|0|0|0
D000001|C0000699|D000515|C0002191|2|0|2004|MBD|0|0|2|ZN|0|2|0|0
D000001|C0000699|D000515|C0002191|1|0|2011|MED|0|0|1|ZN|0|0|0|0
```

Figure 5 - Sample summary_CoOccurs_2013.txt Data

Figure 5 shows that in all of the 2013 MEDLINE Baseline, we have a total of seven occurrences of the D000001/D000515 pair which are found in six different years and all three of our TimeFrames/SABs. We have one occurrence of the pair each year except 2004 where we have two occurrences. Only once were the two MeSH Descriptors marked as Major Topics together (1995) denoted by the ZY.

The **summary_CoOccurs_2013.txt** file is the planned replacement for the current **UMLS MRCOC.RRF** file and is eventually planned to be distributed in flat file, XML, and RDF formats. The first release is only a flat file release. The summary_CoOccurs_2013.txt file is 12 GB uncompressed and 1.3 GB compressed. To replicate the summary results from the *L* entries in the original MRCOC file, you would need to restrict to just the entries with "ZY" for both Major Topics: ***grep D003731 detailed_CoOccurs_2013.txt | grep D006801 | grep "\|ZY\|"***. You can also use the CUIs instead for this command.

If MeSH Qualifier information is required for a given co-occurrence pairing, that information can be derived from the **detailed_CoOccurs_2013.txt** file.

Example

The following example illustrates how the various files and summaries are created using examples from actual MEDLINE citations. Figure 6 displays the indexing from two MEDLINE citations (PMID 20989436 and 16094961) picked to show different aspects of the summarization process. You can see where the initial file Indexing.txt receives a single line for each indexing entry in the citation and that we track all of the relevant information. The highlighted numbers (**[1]**) represent the numbering of each MeSH Descriptor from the two citations and are used to show the co-occurrence creation. The first example is for Version 1 of PMID 20989436 where the earliest date and Publication Date are 19460101 (used default of 1 for month and day since we only had the year), there was no Article Date, Date Completed is 20110128, and the MeSH Indexing year was computed to be 2011. The MeSH Descriptor *Poisoning* **[1]** is a Major Topic, has CUI "C0032343", DUI "D011041", and has no MeSH Qualifiers assigned. The MeSH Descriptor *Poisons* **[2]** is a Major Topic, has CUI "C0032346", DUI "D011042", and has no MeSH Qualifiers assigned. MeSH Descriptor *Vanillic Acid* **[3]** is not a Major Topic, has CUI "C0042315", DUI "D014641", and has one MeSH Qualifier *analogs & derivatives* (AA/Q000031) assigned which is not a Major Topic. The last MeSH Descriptor *Veratrum* **[4]** is an inherited Major Topic, has CUI "C0042527", DUI "D014703", and has one MeSH Qualifier *metabolism* (ME/Q000378) assigned which is a Major Topic.

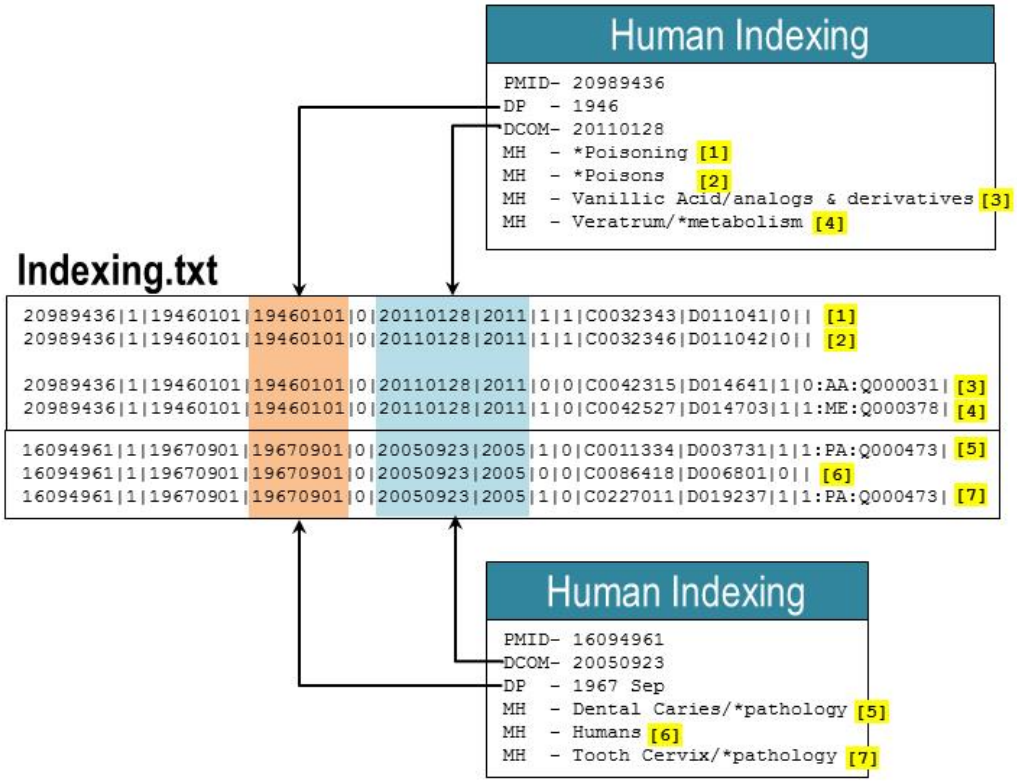

Figure 6 - Example: Human Indexing to Initial Data File

The second example is for Version 1 of PMID 16094961 where the earliest date and Publication Date are 19670901 (used default of 1 for day since we only had the year and month), there was no Article Date, Date Completed is 20050923, and the MeSH Indexing year was computed to be 2005. The MeSH Descriptor *Dental Caries* **[5]** is an inherited Major Topic, has CUI "C0011334", DUI "D003731", and has one MeSH Qualifier *pathology* (PA/Q000473) assigned which is a Major Topic. MeSH Descriptor Humans **[6]** is not a Major Topic, has CUI "C0086418", DUI "D006801", and has no MeSH Qualifiers assigned. The last MeSH Descriptor *Tooth Cervix* **[7]** is an inherited Major Topic, has CUI "C0227011", DUI "D019237", and has one MeSH Qualifier *pathology* (PA/Q000473) assigned which is a Major Topic.

Figure 7 shows all of the co-occurrences derived from the two citations in Figure 6. To make the data more readable, we have separated out the first 7 fields in Figure 7. These 7 fields are on each line in the actual file. We have 1:2, 1:3, 1:4, 2:3, 2:4, and 3:4 from the first citation and 5:6, 5:7, and 6:7 from the second citation. We track whether or not both MeSH Descriptors are Major Topics (ZY) or not (ZN) and track the list of MeSH Qualifiers (if any) attached to each of the Descriptors in the co-occurrence. For example, look at the line ending with **[3,4]** in Figure 7, "*Vanillic Acid/analogs & derivatives* **[3]** co-occurring with

*Veratrum/*metabolism* **[4]** in our Figure 6 example. We end up with ZN since both are not considered Major Topics (only *Veratrum* is) and we have AA (*analogs & derivatives*) assigned to the first Descriptor "*Vanillic Acid*" (D014641) and ME (*metabolism*) assigned to the second Descriptor "*Veratrum*" (D014703).

detailed_CoOccurs_2013.txt

```
20989436|1|19460101|19460101|0|20110128|2011|  (same first 7 fields separated for readability)
     ZY|D011041|1|1|C0032343|0||D011042|1|1|C0032346|0||  [1,2]
     ZN|D011041|1|1|C0032343|0||D014641|0|0|C0042315|1|0:AA:Q000031|  [1,3]
     ZY|D011041|1|1|C0032343|0||D014703|1|0|C0042527|1|1:ME:Q000378|  [1,4]
     ZN|D011042|1|1|C0032346|0||D014641|0|0|C0042315|1|0:AA:Q000031|  [2,3]
     ZY|D011042|1|1|C0032346|0||D014703|1|0|C0042527|1|1:ME:Q000378|  [2,4]
     ZN|D014641|0|0|C0042315|1|0:AA:Q000031|D014703|1|0|C0042527|1|1:ME:Q000378|  [3,4]

16094961|1|19670901|19670901|0|20050923|2005|  (same first 7 fields separated for readability)
     ZN|D003731|1|0|C0011334|1|1:PA:Q000473|D006801|0|0|C0086418|0||  [5,6]
     ZY|D003731|1|0|C0011334|1|1:PA:Q000473|D019237|1|0|C0227011|1|1:PA:Q000473|  [5,7]
     ZN|D006801|0|0|C0086418|0||D019237|1|0|C0227011|1|1:PA:Q000473|  [6,7]
```

Figure 7 - Detailed Co-Occurrence for Example Citations

Figure 8 shows the final summarized results categorized by the Date Completed Years represented in our example citations. We are only including years where the overall frequency is at least one, so if a co-occurrence does not show up in a given year, that year will not be shown. The lines in Figure 8 illustrate this point with only entries for the Date Completed Years from the two citations (2005 and 2011). If we revisit our *Vanillic Acid/analogs & derivatives* and *Veratrum/*metabolism* co-occurrence, we can see the summary is the very last line in Figure 8. Reading left to right, we have MeSH Descriptor *Vanillic Acid* (D014641|C0042315) co-occurring with MeSH Descriptor *Veratrum* (D014703|C0042527) with frequency of one overall and zero as both being marked as Major Topics in 2011. 2011 falls into the **MED** TimeFrame/SAB category, both MeSH Descriptors have MeSH Qualifiers in our example, so the Frequency No Qualifiers is zero. Only one of the MeSH Descriptors is a Major Topic and both have MeSH Qualifiers assigned, so Frequency Starred No Qualifier and Frequency Starred Qualifiers are also zero. Both MeSH Descriptors are not designated as Major Topics, so the ZN flag is used. The first MeSH Descriptor *Vanillic Acid* is not a Major Topic, so Frequency DUI1 Only Starred is zero. The second MeSH Descriptor *Veratrum* is a Major Topic and the first MeSH Descriptor is not, so Frequency DUI2 Only Starred frequency is one. Both MeSH Descriptors have MeSH Qualifiers assigned, so the last two fields Frequency DUI1 Only has Qualifier and Frequency DUI2 Only has Qualifier are both zero.

summary_CoOccurs_2013.txt

```
D003731|C0011334|D006801|C0086418|1|0|2005|MBD|0|0|0|ZN|1|0|1|0  [5,6]
D003731|C0011334|D019237|C0227011|1|1|2005|MBD|0|0|1|ZY|0|0|0|0  [5,7]
D006801|C0086418|D019237|C0227011|1|0|2005|MBD|0|0|0|ZN|0|1|0|1  [6,7]
D011041|C0032343|D011042|C0032346|1|1|2011|MED|1|1|0|ZY|0|0|0|0  [1,2]
D011041|C0032343|D014641|C0042315|1|0|2011|MED|0|0|0|ZN|1|0|0|1  [1,3]
D011041|C0032343|D014703|C0042527|1|1|2011|MED|0|0|0|ZY|0|0|0|1  [1,4]
D011042|C0032346|D014641|C0042315|1|0|2011|MED|0|0|0|ZN|1|0|0|1  [2,3]
D011042|C0032346|D014703|C0042527|1|1|2011|MED|0|0|0|ZY|0|0|0|1  [2,4]
D014641|C0042315|D014703|C0042527|1|0|2011|MED|0|0|0|ZN|0|1|0|0  [3,4]
```

Figure 8 - Final Co-Occurrence Summary for Example Citations

Appendix I: Dates

The new MRCOC file contains four different dates where they are available including: Publication Date, Article Date, Date Completed, and MeSH Indexing Year. We also provide a fifth field with the calculated earliest date between PubDate, Article Date, and Date Completed.

Publication Date / Date of Publication / DP / PubDate[10]:

Publication date contains the full date on which the issue of the journal was published. The standardized format consists of elements for a 4-digit year, a 3-character abbreviated month, and a 1 or 2-digit day. Every record does not contain all of these elements; the data are taken as they are published in the journal issue, with minor alterations by NLM such as abbreviating months. Journals vary in the way the publication date appears on an issue. Some journals include just the year, whereas others include the year plus month or year plus month plus day. And, some journals use the year and season (e.g., Winter 1997). The publication date in the citation is recorded as it appears in the journal.

XML Tag: <PubDate>
Potential XML Tags:
<Year>
<Month>
<Day>
<Season>
<MedlineDate>

The Publication Date is the most complex of the dates and we have assigned defaults in some cases so we could compute a consistent YYYYMMDD format for the Publication Date. These defaults are noted below:

For Seasons (<Season> and <MedlineDate>), we will use the following month and days:
Spring (spr) - March 20
Summer (sum) - June 21
Fall/Autumn - September 22
Winter - December 21
Christmas - December 25
Easter - March 20

For Semesters, Trimesters, and Quarters (<MedlineDate>, we will use the following month and days:
1st - January 1
2nd - April 1
3rd - July 1
4th - October 1

In the event that either the month or the day is missing or not derivable, the default is "1". For example, if an entry has Year 1967, but, no Month or Day specified, we will use the defaults and assign January 1, 1967.

[10] http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html#pubdate

Article Date / Date of Electronic Publication / DEP / ArticleDate[11]:

Article Date contains the date the publisher made an electronic version of the article available, with the month represented as a 2-digit numeric rather than an alphabetic abbreviation as is the case for the month in PubDate. A record includes <ArticleDate> only if that data are included in the publisher's electronic submission to NLM, and it may be present on records with <Article> PubModel attribute values of Electronic, Print-Electronic, Electronic-Print, or Electronic-eCollection. Article Date is well behaved when present – it will always contain Year, Month, and Day.

XML Tag: <ArticleDate DataType="">
Potential XML Tags:
<Year>
<Month>
<Day>

Date Completed / DCOM / DateCompleted[12]:

Date Completed is the date processing of the record ends; i.e., MeSH Headings have been added, quality assurance validations are completed, and the completed record subsequently is distributed to PubMed and licensees. For records in the OLDMEDLINE subset: <DateCompleted> is the approximate date the record entered PubMed instead of the date processing ends because OLDMEDLINE records are created and processed differently from MEDLINE records. Completion Dates are well behaved and always contain the Year, Month, and Day tags.

XML Tag: <DateCompleted>
XML Tags:
<Year>
<Month>
<Day>

MeSH Indexing Year[13]:

The MeSH Indexing Year typically runs from mid-November to mid-November the following year. This refers to the version of MeSH used by the NLM Indexers to manually index citations during that time frame. For example, MeSH 2012 refers to the 2012 MeSH Indexing Year and includes Date Completed dates from November 18, 2011 through November 13, 2012. The inclusion of the MeSH Indexing Year is helping provide context for the co-occurrences. We currently have MeSH Indexing Year date ranges for 1964 to the Present. Any Date Completed dates prior to the 1964 MeSH Indexing Year will have their MeSH Indexing Year calculated using November 20 as the changeover date. For a complete and up-to-date list of the MeSH Indexing Years, please see the following URL: ***http://www.nlm.nih.gov/bsd/mesh_indexing_date_range.html***

[11] http://www.nlm.nih.gov/bsd/licensee/elements_article_source.html
[12] http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html#datecompleted
[13] http://www.nlm.nih.gov/bsd/mesh_indexing_date_range.html